Bagging and Other Ensemble Methods

Sargur N. Srihari
srihari@buffalo.edu

Regularization Strategies

1. Parameter Norm Penalties
2. Norm Penalties as Constrained Optimization
3. Regularization and Under-constrained Problems
4. Data Set Augmentation
5. Noise Robustness
6. Semi-supervised learning
7. Multi-task learning
8. Early Stopping
9. Parameter tying and parameter sharing
10. Sparse representations
11. Bagging and other ensemble methods
12. Dropout
13. Adversarial training
14. Tangent methods

What is bagging?

- It is short for *Bootstrap Aggregating*
- It is a technique for reducing generalization error by combining several models
  - Idea is to train several models separately, then have all the models vote on the output for test examples
- This strategy is called *model averaging*
- Techniques employing this strategy are known as *ensemble methods*
- Model averaging works because different models will not make the same mistake

Ex: Ensemble error rate

- Consider set of k regression models
  - Each model makes error ε_i on each example, $i=1,..N$
  - Errors drawn from a zero-mean multivariate normal with variance $\mathrm{E}[\varepsilon_i^2]=v$ and covariance $\mathrm{E}[\varepsilon_i\varepsilon_j]=c$
    - Error of average prediction of all ensemble models: $\boxed{\frac{1}{k}\sum_i \varepsilon_i}$
  - Expected squared error of ensemble prediction is

$$\boxed{E\left[\left(\frac{1}{k}\sum_i \varepsilon_i\right)^2\right] = \frac{1}{k^2}E\left[\sum_i\left(\varepsilon_i^2 + \sum_{j\neq i}\varepsilon_i\varepsilon_j\right)\right] = \frac{1}{k}v + \frac{k-1}{k}c}$$

    - If errors are perfectly correlated, $c=v$, and mean squared error reduces to v, so model averaging does not help
    - If errors are perfectly uncorrelated and $c=0$, expected squared error of ensemble is only v/k
      - Ensemble error decreases linearly with ensemble size

Ensemble vs Bagging

- Different ensemble methods construct the ensemble of models in different ways
  - Ex: each member of ensemble could be formed by training a completely different kind of model using a different algorithm or objective function
- Bagging is a method that allows the same kind of model, training algorithm and objective function to be reused several times

The Bagging Technique

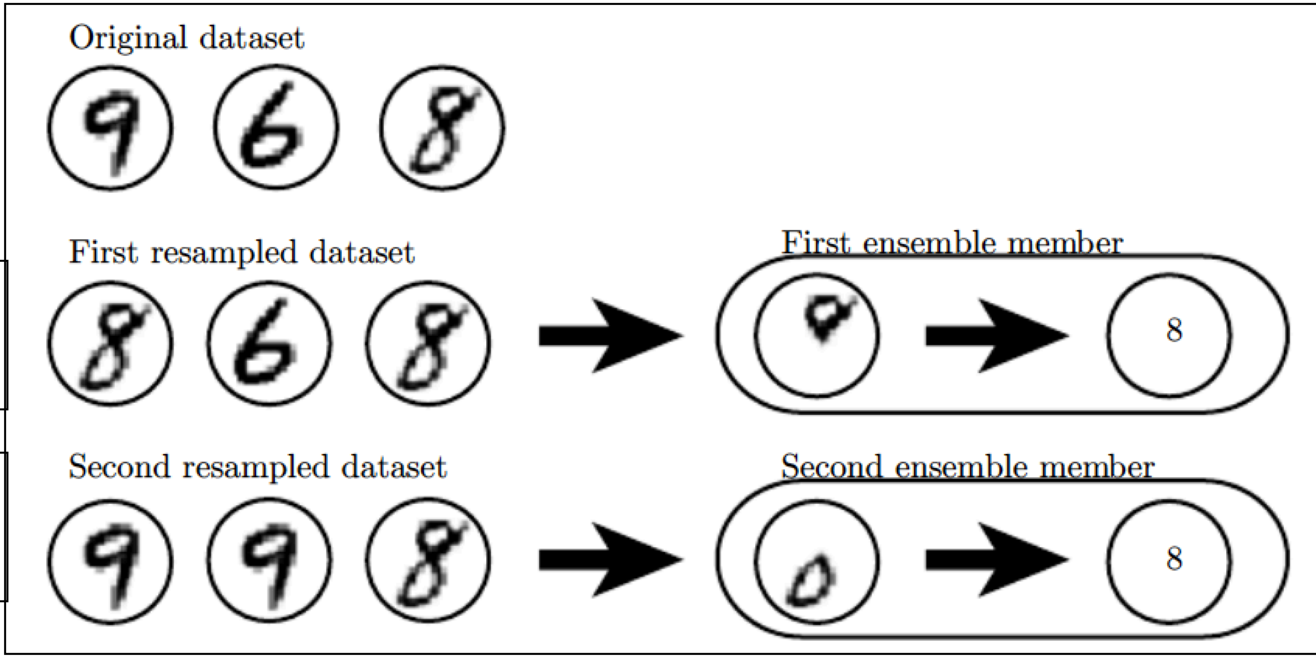
- Given training set D of size N, generate k data sets of same no of examples as original by sampling with replacement
  - Some observations may be repeated in each D_i the rest being duplicates. This is known as a bootstrap sample
    - The differences in examples will result in differences between trained models
  - The k models are combined by averaging the output (for regression) or voting (for classification)
- An example is given next

Example of Bagging Principle

- Task of training an 8 detector
- Bagging training procedure
  - make different data sets by resampling the given data set

Omit 9
Repeat 8

Omit 6
Repeat 9



Detector learns that loop on top Is an 8

Detector learns that loop at bottom is an 8

- Each detector is brittle
- Their average is robust achieving maximum confidence when both loops are present

Neural nets and bagging

- Neural nets reach a wide variety of solution points
  - Thus they benefit from model averaging when trained on the same dataset
  - Differences in:
    - random initializations
    - random selection of minibatches, in hyperparameters,
  - cause different members of the ensemble to make partially independent errors

Model averaging is powerful

- Model averaging is a reliable method for reducing generalization error
  - Machine learning contests are usually won by model averaging over dozens of models
    - Ex: Netflix grand prize
- Since model averaging performance comes at the expense of increased computation and memory, benchmark comparisons are made using a single model

Boosting

- Incrementally adding models to the ensemble
  - After a weak learner is added, the data are reweighted:
    - examples that are misclassified gain weight and examples that are classified correctly lose weight
- Has been applied to ensembles of neural networks, by incrementally adding neural netowrks to the ensemble
- Also interpreting a neural network as an ensemble, incrementally adding hidden units to the network